Supplementary Materials of Speech2Lip: High-fidelity Speech to Lip Generation by Learning from a Short Video

S1. Implementation Details

S1.1. Network Details

Implicit model. The implicit model $\boldsymbol{f_\theta}$ is defined as an 8-layer multi-layer perceptron (MLP) with 256 channels, with each layer accompanied by a ReLU activation function. It is worth noting that the inputs (*i.e.*, $\boldsymbol{x_{c,n}}$ and $\boldsymbol{t}_s$) are actually the results of the Positional Encoding [7] of the original data. The use of Positional Encoding allows for the exploitation of high-frequency information, resulting in the synthesis of high-resolution output. In detail, it encodes a set of functions to represent an arbitrary input data p, as

$$\gamma(p) = [\sin(\pi p), \cos(\pi p), \cdots, \sin(2^{L-1}\pi p), \cos(2^{L-1}\pi p)]. \tag{1}$$

We empirically set $L = 10$ as [7].

Blend-Net. Blend-Net predicts a residual map of pixels in order to effectively eliminate any artifacts that may have resulted from the paste operation. In the process of designing Blend-Net, we have taken great care to adopt a UNet-like structure that is simplistic in nature, yet powerful in its ability to extract features of varying scales. It is noteworthy to mention that our encoder and decoder frameworks are composed of six and five carefully crafted CNN layers, respectively. These details are highlighted in Figure S1.

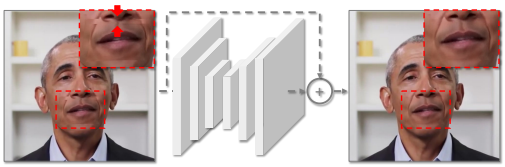

Figure S1. Structure of Blend-Net.

S1.2. Training Details

The model is implemented in PyTorch [8] with Adam optimizer [4], and the learning rate is set to be 1e-4. Our model can be trained using one NVIDIA GeForce RTX 3090. It is worth noting that the sync network requires a well-initialized lip image, and therefore, end-to-end training can result in collapse. To avoid this, we first train the model for 100k iterations using ω_m=1.0, ω_w=1.0, ω_d=1.0 and ω_s=0.0 (details are described in Eq. 10 in the main paper) to learn satisfactory visual quality. Then, for the subsequent 200k iterations, we add $\mathcal{L}_s$ to improve audio synchronization, with ω_s being set to 1.

S1.3. Data Preprocessing

Ground-truth lip images. The offline preprocessing of ground-truth lip images $I^m_{o\to c}$ is depicted in Figure 4 of the main paper. Based on our empirical experiments, the lip boundary is accurately determined (as indicated by the orange bounding box). The accompanying motion heatmaps also demonstrate that only a small area is affected by speech. In cases where the expected mouth motion is exceptionally large, the output may appear unnatural; however, such an occurrence is rare, based on our data analysis.

Audio preprocessing. To preprocess the audio data, the raw audio signal is initially fed into a pre-trained DeepSpeech network [1]. Then, it undergoes temporal smoothing via a 1D convolutional network as AD-NeRF [3].

S2. Experiments

S2.1. Comparisons on More Datasets.

Table S1 presents an evaluation of our **Speech2Lip** across two datasets, following [3, 6]. It should be noted that LSP [6] has also provided two additional training videos, which feature the subject Nadella and the subject Obama. However, we do not compare our model's performance on these two videos due to multiple shot changes with varying focal lengths, which has made it difficult to identify a 3-5 minute long video clip for training and evaluation. As our problem setting pertains to speaker-specific talking heads, we focus on comparisons with other models under this setting. Specifically, Testset IV (consisting of subject Obama, with a resolution of 450×450) and Testset V (consisting of subject McStay, with a resolution of 550×550) are gathered from [3] and [6], respectively. Since [6] did not provide pre-trained models for Testset IV, we have left this field blank in the table. Our algorithm outperforms the others, as evidenced by the quantitative results.

Method	Trained with large extra data	Testset IV					Testset V				
		PSNR↑	SSIM↑	CPBD↑	LMD↓	Sync↑	PSNR↑	SSIM↑	CPBD↑	LMD↓	Sync↑
Ground Truth	*N/A*	*N/A*	*1.000*	*0.260*	*0.000*	*8.483*	*N/A*	*1.000*	*0.320*	*0.000*	*8.556*
LSP [6]	No	-	-	-	-	-	29.076	0.726	0.243	3.830	5.770
AD-NeRF [3]	No	32.684	0.931	0.231	**3.092**	3.722	28.858	0.812	0.308	3.810	4.981
DFRF [9]	No	33.385	0.967	0.239	3.095	3.831	30.259	0.918	0.313	4.016	4.127
Speech2Lip	No	**33.791**	**0.971**	**0.258**	3.701	**4.527**	**31.737**	**0.921**	**0.318**	**3.767**	**6.884**

Table S1. Quantitative results compared with the previous SOTA methods. Image quality assessment metrics (*i.e.*, PSNR, SSIM, and CPBD) are computed within **mouth region**. The best results are **bold**.

S2.2. Comparisons with More Speaker-specific Models

We further compare our algorithm against speaker-specific models, such as SynObama [10], NVP [11], and SSP-NeRF [5], which have released neither code nor pre-trained models, but demo videos. Specifically, to synthesize images, we extract speech from the demo of SSP-NeRF [5], which features video clips for comparison with both SynObama [10] and NVP [11]. We utilize Testset VI for comparing with SynObama [10] and SSP-NeRF [5], and Testset VII for comparing with NVP [11] and SSP-NeRF [5]. Qualitative results are provided in Figure S2. Given the absence of any ground-truth images, we emphasize the corresponding letters and the results demonstrate the superiority of our algorithm.

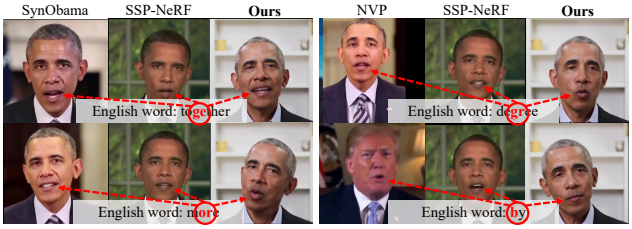

Figure S2. Qualitative results with corresponding letters being highlighted.

S3. More Analyses

S3.1. Empirical Study and Motivations

We select some video clips from our dataset that a speaker is pronouncing the same word ("**po**ur" and "offi**ci**er" in French). As seen in Figure S3, the subjects display a range of head poses and appearances, indicating their insensitivity to the input speech. Relying solely on speech to generate the synthesis of all facial areas can result in ambiguous training signals, which may force the model to learn inaccurate correlations, particularly when the training data is limited. This highlights the need for a more targeted approach in modeling speech-synchronized facial animation.

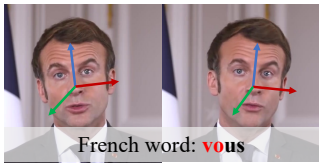

Figure S3. Different head motions with similar speeches.

S3.2. Analysis about Motion Heatmap

In line with Section 3 of the main paper, we provide additional examples of motion heatmap to showcase the key insight of our method: the disentanglement of speech-sensitive and speech-insensitive appearance and motion. This facilitates the synthesis of high-fidelity results.

The motion heatmap is defined as the variance map of the image sequence. Figure S4 depicts the motion heatmaps (the last column) for images captured under observed views (the first column) that are warped to the canonical space (the second column) based on 3DMM [2]).

In Figure 3 of the main paper, it has been established that head motion is insensitive to input speech. However, upon the elimination of head motions, the major motion is concentrated around the lip region (as depicted in the last column of Figure S4). In such a scenario, a simple yet efficient solution is to learn a mapping from input speech to the lip area. This stimulates the devisal of our proposed decomposing-synthesis-composition framework, which disentangles speech-sensitive and speech-insensitive appearance and motion, enabling our model to learn more effectively from short data.

S3.3. Analysis about Sync Score

To verify the relationship between the Sync score and synchronization quality, we devised a toy experiment that involved the generation of unsynced videos through the shifting of image sequences. As shifting offsets become larger, the synchronization quality becomes worse, therefore we can evaluate corresponding Sync scores based on the synchronization quality. As demonstrated in Figure S5, the Sync score is sensitive to synchronization quality only within a certain range.

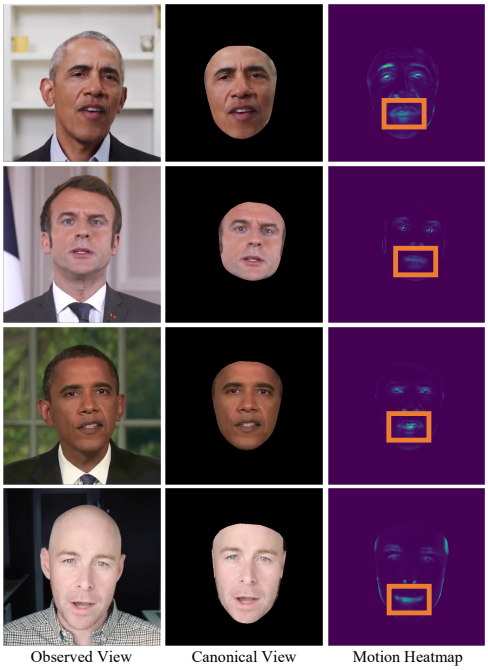

Figure S4. More speech-sensitive motion heatmaps.

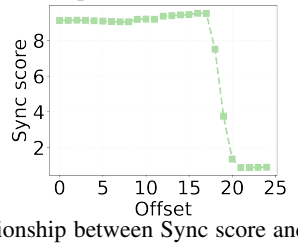

Figure S5. Relationship between Sync score and synchronization quality.

S4. Limitations and Discussions

While the proposed decomposing-synthesis-composition framework can generate reasonably good results using a very short video ($3 \sim 5$ minutes), we also inherit the same limitation from the speaker-specific setting. Specifically, this means that 1) individual models must be trained for each particular person and 2) the range of appearances displayed by the generated lip sequences is constrained by the available training data.

Our proposed framework can synthesize realistic, high-fidelity talking head videos using only very little training data, which may be used to manipulate media, such as videos or images, that can be used to deceive or spread disinformation. It may have potentially negative implications for society, including political manipulation, financial fraud, and reputational damage. Therefore we urge caution to prevent any potential improper use.

In the future, we plan to make the model adaptive to environmental changes (*e.g.* illumination changes), which can be valuable for practical applications. We also aspire to enhance the model's ability to generalize and refine its output quality while decreasing the amount of necessary training data.

References

[1] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016. 1

[2] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 2

[3] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021. 1, 2

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[5] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. *arXiv preprint arXiv:2201.07786*, 2022. 2

[6] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG)*, 40(6):1–17, 2021. 1, 2

[7] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 1

[8] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 1

[9] Shuai Shen, Wanhua Li, Zheng Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. In *European conference on computer vision*, 2022. 2

[10] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 2

[11] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *European conference on computer vision*, pages 716–731. Springer, 2020. 2